Bayesian Hierarchical Bias Model for Demonstrating Biosimilarity

Joseph Wu[*] Sandeep Menon[†] Gheorghe Doros[‡] Kerry Barker[§]
Mark Chang[¶]

Abstract
In 2010, the passage of the Biologics Price Competition and Innovation Act (BPCI) created an abbreviated licensure pathway in section 351(k) of the Public Health Service Act (PHS). This new law allows for an expeditious approval process for a generic follow-on biological product shown to be biosimilar to a licensed reference biological product. Traditional statistical methods used to test for average bioequivalence as in a generic drug development may not be the most efficient way to apply to biosimilarity. We adopt a Bayesian approach to establish biosimilarity for a composite endpoint. Specifically, we propose a hierarchical bias model to capture the effect difference between the reference and follow-on products. Within a non-inferiority framework, we formulate a statistical test using the posterior distributions to demonstrate biosimilarity. We illustrate this proposed methodology using a recombinant polypeptide example used to treat rheumatoid arthritis and the composite endpoint of ACR20. Using simulation, we have shown that the type I error is preserved when reference product is not performing worse in current trial than historical trial. Statistical power is better than the frequentist approach as sample size increases.

Key Words: Biosimilarity, Follow-on biologics, Non-inferiority, Bayesian inference, Composite endpoint, Rheumatoid arthritis.

1. Introduction

The concept of biosimilarity has received increasing popularity within the scientific community recently. One big motivation to explore biosimilar products is the unprecedented opportunity gradually opened up by numerous soon-to-be expiring licenses of major biological products. In 2010, the passage of the Biologics Price Competition and Innovation Act (BPCI) created an abbreviated licensure pathway in section 351(k) of the Public Health Service Act (PHS). This new law allows for an expeditious approval process for a generic follow-on biological product shown to be biosimilar to a licensed reference biological product. Section 351(i) of the PHS Act defines biosimilarity to mean "that the biological product is highly similar to the reference product notwithstanding minor differences in clinically inactive components" and that "there are no clinically meaningful differences between the biological product and the reference product in terms of the safety, purity, and potency of the product." Due to their large and complex molecular structures, biological products are fundamentally disparate from small synthetic drugs, and so are their mechanisms of action. Traditional statistical methods used to test for bioequivalence as in a generic drug development may not be the most efficient way to apply to biosimilarity (Kang and Chow, 2012). Many of the recently proposed methods to establish biosimilarity between an innovator reference biological product and a generic follow-on biological product primarily borrowed

[*] Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118
[†] Pfizer, Inc., 300 Technology Square, Cambridge, MA 02139
[‡] Department of Biostatistics, Boston University School of Public Health, 801 Massachusetts Avenue, Boston, MA 02118
[§] Pfizer, Inc., 10 Fawcett Street, Suite 203, Cambridge, MA 02138
[¶] AMAG Pharmaceuticals, 1100 Winter Street, Waltham, MA 02451

ideas from average bioequivalence (ABE) trials. An ABE trial using a 2 × 2 crossover design is the standard approach suggested by the U.S. Food and Drug Administration (FDA) to test for the equivalence between a reference drug and a new generic drug.

Biological products are fundamentally different from small-molecule compound. They are large polypeptide molecules with a much larger molecular weight than small-molecule synthetic drugs. Therefore, they tend to have a longer half-life and require a longer washout period. In this case, the standard crossover design normally used for bioequivalence trial may not be efficient if applied to biosimilarity trials. A more appropriate design would be the parallel group design. Various biosimilarity criteria have been suggested and they depend on the study designs and objectives. For a parallel three-arm trial with two of the arms for the reference product from two different manufacturing lots and the other one for the follow-on biological product, Kang and Chow (2012) proposed the relative distance as a biosimilarity criterion. The authors developed a test that assumes asymptotic distribution of its maximum likelihood estimator (MLE). Lin *et al.* (2012) presented the parallel line assay design that requires two dose-response trials for both the reference and follow-on biological products. Under the assumption of the parallel line bioassay, they assumed a linear relationship between the binary efficacy endpoint and the dose-dependent mean product characteristic. The biosimilarity criterion in this case is the relative potency.

This paper is motivated by the need to develop an innovative statistical method for proving biosimilarity. Here is the organization of the subsequent sections. Section 2 introduces the composite endpoint of interest and describes the proposed clinical study for the demonstration of biosimilarity that uses this composite endpoint. This section also provides the rationale for a non-inferiority (NI) testing framework and a Bayesian inferential approach to achieve the study's objectives. Section 3 describes the details of a simulation plan to examine the operating characteristics of this proposed method and also summarizes the simulation results with comparison to the frequentist approach. Section 4 discusses the overall results and proposes further work in this area.

2. Biosimilarity Using Composite Endpoint

Although in the past few years some statistical methods have been proposed for the case of a single primary efficacy endpoint, some biological products are designed to treat medical conditions with improvement measured by several endpoints. For example, rheumatoid arthritis (RA) is a disease of the immune system that leads to the inflammation in the joints. In clinical trials studying RA, the current standard measure of efficacy is the ACR20 criteria recommended by the American College of Rheumatology (ACR) Committee. For each individual patient in a trial, it measures if this patient has experienced a clinical response of overall improvement by evaluating the percentage of improvement in a core set of variables during the trial. Generally speaking, if a patient experiences at least 20% improvement from baseline in multiple variables simultaneously, this patient is defined as having satisfied the definition of a clinical response. Therefore, ACR20 is a composite criterion and has served as a working model to other disorders that currently require multiple primary endpoints (Offen *et al.*, 2007). The percent change in each of these variables is also measured at different time points such as 3, 6 and 12 months and one of the time points is used to establish primary efficacy. Table 1 summarizes the ACR20 improvement criteria (Felson, Anderson, Boers *et al.*, 1993).

Table 1: ACR20 Improvement Criteria

Quantitative criterion	Endpoints
percent reduction	1. Tender joint count, and
$\geqslant 20\%$ improvement in	2. Swollen joint count, and
	At least 3 of the following:
	3. Physician global assessment of disease activity
	4. Patient global assessment of disease activity
	5. Patient assessment of pain (e.g. Visual Analog Scale)
	6. Physical disability or functionality
	7. Inflammatory marker: erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP)

2.1 Non-Inferiority Hypothesis

Motivated by the need to decrease sample size for a clinical study, we propose a non-inferiority framework to test for biosimilarity using the clinical data. This non-inferiority trial design allows the current biosimilarity trial to meaningfully connect to any similarly conducted historical trials that have evaluated the effect of the licensed reference biological product. Since a standard treatment is already available for the medical condition, including a placebo arm in the current trial will not be ethical. We can use k to index the biological product with $k = 1$ representing the innovator reference product and $k = 2$ the proposed follow-on biological product. In this case, we are interested in testing if, based on the composite endpoint, the proposed biological product is not inferior to the licensed biological product.

In this two-arm design, patients are randomized to either the original reference or the follow-on generic biological product. For each patient, outcomes on J multiple endpoints will be measured at pre-specified follow-up times. These J endpoints can be generally considered as independent measures. We can use x_{kji} to denote the jth endpoint $(j = 1, 2, ..., J)$ observed in the ith patient receiving the product k. In this case, we can assume that it is normally distributed as

$$x_{kji} \sim N(\mu_{kj}, \sigma_k^2) \tag{1}$$

where $k = 1$ or 2, and $i = 1, 2, ..., n_k$. The fixed randomization ratio is therefore equal to $R = n_2/n_1$. μ_{kj} is the mean response for the jth endpoint and σ_k^2 is the variance which is assumed to be the same for all J endpoints but different between the products. In addition, we want to consider combining these J endpoints into a single composite binary efficacy endpoint y_{ki} which can be generally defined as

$$y_{ki} \quad = \quad \begin{cases} 1 & \boldsymbol{x}_{ki} \geqslant \boldsymbol{\omega} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $\boldsymbol{x}_{ki} = (x_{k1i}, x_{k2i}, ..., x_{kJi})'$ is the random vector of outcomes for the ith patient and $\boldsymbol{\omega} = (\omega_1, \omega_2, ..., \omega_J)'$ is a J-dimensional vector of cutoff points for the endpoints common to both biological products, assuming that higher values of x_{kji}'s are desirable. If we denote the probability of a response on the composite endpoint for product k as p_k, then $p_k = P(y_{ki} = 1) = P(\boldsymbol{x}_{ki} \geqslant \boldsymbol{\omega})$, and our non-inferiority hypotheses of interest can be constructed as

$$H_0 : p_2 - p_1 \leqslant -\delta \qquad \text{versus} \qquad H_A : p_2 - p_1 > -\delta. \tag{3}$$

where $\delta(\delta > 0)$ is the pre-specified non-inferiority margin for the difference between the two probabilities.

2.2 Bayesian Hierarchical Bias Model

Since multiple endpoints are considered in developing the composite endpoint, the Bayesian approach allows for the borrowing of estimative strength between the J multiple endpoints on the precision parameters in addition to the borrowing between historical trials. This also means that fewer subjects may be needed for the reference product and more subjects can be randomized to the new and potentially biosimilar product. This is a realization of the FDA guidance regarding its suggestion to use smaller clinical studies and to convene them based on results from previously conducted studies. Furthermore, the composite endpoint can be defined by criteria on the multiple endpoints which provide clinically meaningful interpretation. According to (2), p_k will be defined as a function of the parameters such that $p_k = f(\mu_{k1}, \mu_{k2}, ..., \mu_{kJ}, \sigma_k^2)$ for $k = 1$ or 2.

In this proposed hierarchical bias model, we allow the inclusion of any number of historical trials for the licensed reference product. For example, if there are H historical trials available before the conduct of the current biosimilarity trial, we can let x_{1hji} be the value of the jth endpoint observed for the ith patient receiving the original reference product $k = 1$ in the hth historical trial such that

$$x_{1hji} \sim N(\mu_{1hj}, \sigma_1^2) \tag{4}$$

where $i = 1, 2, ..., n_{1h}$, $h = 1, 2, ..., H$, and $j = 1, 2, ..., J$. In the above model, we assume that these H trials and the current biosimilarity trial share the same within-study variance parameter σ_1^2 and it is also assumed to be constant across all J endpoints. This assumption allows borrowing between the historical trials and also between the J endpoints. In addition, we can represent the jth sample mean as $\overline{x}_{1hj}$ which is equal to $(\sum_{i=1}^{n_{1h}} x_{1hji})/n_{1h}$. Other sample means can be similarly defined.

Additionally, under exchangeability, we consider the mean parameters, μ_{1j} of the current biosimilarity trial and μ_{1hj} of the hth historical trial, for the original reference product, come from the same distribution as

$$\mu_{1j}, \mu_{1hj} \sim N(\mu_{1j}^o, \sigma_{1b}^2) \tag{5}$$

where $h = 1, 2, ..., H$. μ_{1j}^o is the overall mean and σ_{1b}^2 is the between-trial variance parameter, which is assumed to be the same across the J endpoints. Hierarchical modeling is a logical way of combining historical data when exchangeability between parameters is highly plausible, and as in the current problem, these historical trials used an efficacy response defined by the same criterion. This hierarchical structure implies heterogeneity of the mean endpoints.

For the new generic follow-on product in the current biosimilarity trial, we think of its mean response on the j endpoint, μ_{2j}, as having a bias term from that of the mean endpoint of the original product, μ_{1j}. Pocock (1976) discussed the parameterizing a bias term to model the difference between mean of historical control and the same control but in the current trial. Therefore, we propose this relationship

$$\mu_{1j} = \mu_{2j} + \xi_j \tag{6}$$

where ξ_j represents the bias of μ_{2j} from μ_{1j}. If ξ_j is equal to 0, then $\mu_{2j} = \mu_{1j}$ meaning that the follow-on product has the same mean as the licensed reference product on the jth endpoint. If $\xi_j < 0$, then it means the follow-on product exhibits a better effect than the reference product in the jth endpoint, and the opposite interpretation follows if $\xi_j > 0$. Since we do not know the true value of ξ_j, we can assume a model for this bias parameter as

$$\xi_j \sim N(\theta, \sigma_\xi^2) \tag{7}$$

where $j = 1, 2, ..., J$. We center the expectation of ξ_j skeptically at the null hypothesis, θ, to allow the data to reflect and influence its true direction and magnitude away from the null value. The null value θ is the margin on the scale of individual endpoints, such that when this margin is uniformly subtracted from all of the mean responses, the probability of the binary composite endpoint will decrease by exactly the amount of δ as in $f(\mu_{k1} - \theta, \mu_{k2} - \theta, ..., \mu_{kJ} - \theta, \sigma_k^2) - f(\mu_{k1}, \mu_{k2}, ..., \mu_{kJ}, \sigma_k^2) = -\delta$. This relationship between δ and θ is one-on-one. Therefore, centering the mean of ξ_j on θ also suggests that μ_{1j} and μ_{2j} are dissimilar to begin with. We also assume that the variance parameter σ_ξ^2 to be the same across all J endpoints but a large σ_ξ^2 will suggest that this distribution is only weakly informative. Figure 1 displays the graphical representation of this model with each circle representing a random node, a single-line arrow representing the dependent stochastic relationship and a double-line arrow representing a logical relationship.

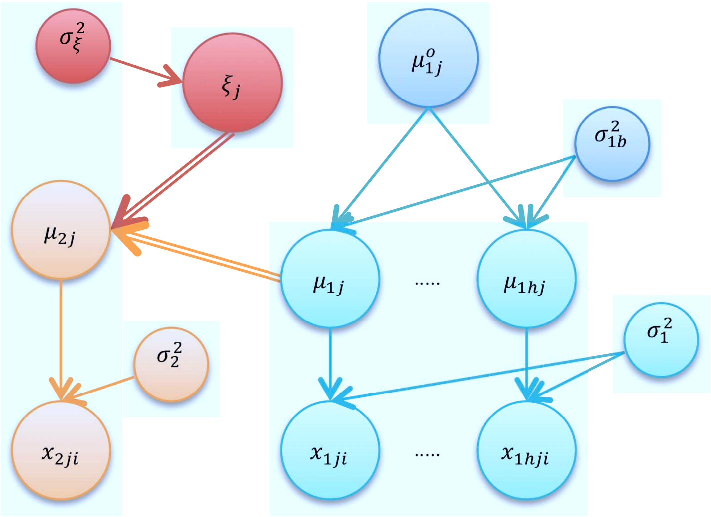

Figure 1: *Graphical representation of the proposed Bayesian hierarchical bias model,* $j = 1, 2, ..., J$.

This hierarchical bias model completely specifies the likelihood function, and we can consider generally uninformative prior distributions such as Jeffery's priors for the parameters. Using Gibbs sampling, we can then directly estimate the posterior probability

$$P(p_2 - p_1 > -\delta | \boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, \boldsymbol{x}_{2j}, j = 1, 2, ..., J)$$

$$= E[I(p_2 - p_1 > -\delta)|\boldsymbol{x}_{1j}, \boldsymbol{x}_{11j}, ..., \boldsymbol{x}_{1Hj}, \boldsymbol{x}_{2j}, j = 1, 2, ..., J].$$

The decision rule is to reject the null hypothesis when this posterior probability is greater than a critical probability p_c which can be pre-specified as high as 95% or 97.5% depending on the clinical significance.

2.3 Determination of Bayesian Non-Inferiority Margin

Another major challenge in a non-inferiority trial design is to determine the NI margin δ and hence its corresponding θ for each of the individual endpoints. One way to specify δ is to mirror the fixed margin method in the frequentist paradigm in the current Bayesian paradigm (Gamalo, Wu, Tiwari, 2012; Gamalo, Tiwari, LaVange, 2013). In the frequentist paradigm, the NI margin is set to be the lower bound of the $100\%(1-\alpha)$ confidence interval for the effect $p_{1h'} - p_{0h'}$ in a selected historical placebo-controlled trial h', where $0h'$ represents the placebo arm and $1h'$ represents the innovator reference product arm in this trial. This historical placebo-controlled trial h' was usually a trial that led to its first FDA approval.

If we assume a similar model as in (4) for the placebo and treatment arms in this placebo-controlled trial

$$x_{kh'ji} \sim N(\mu_{kh'j}, \sigma^2_{kh'}),$$

where $k = 0$ or 1 and elicit a flat non-informative prior for $\mu_{kh'j}$ as in $P(\mu_{kh'j}) \propto 1$ and a Jeffery's prior for the variance $\sigma^2_{kh'}$ as in $P(\sigma^2_{kh'}) \propto 1/\sigma^2_{kh'}$, then

$$\begin{aligned} \mu_{kh'j}|\boldsymbol{x}_{kh'j}, j = 1, 2, ..., J &\sim N\left(\overline{x}_{kh'j}, \frac{\sigma^2_{kh'}}{n_{kh'}}\right) \\ \sigma^2_{kh'}|\boldsymbol{x}_{kh'j}, j = 1, 2, ..., J &\sim IG\left(\frac{Jn_{kh'}}{2}, \frac{1}{2}\sum_{j=1}^{J}\sum_{i=1}^{n_{kh'}}(x_{kh'ji} - \mu_{kh'j})^2\right). \end{aligned} \tag{8}$$

We can use Gibbs sampling to simulate for $p_{1h'} - p_{0h'}$ and solve for δ as the lower bound of the $100\%(1-\alpha)$ credibility interval such that

$$P(p_{1h'} - p_{0h'} > \delta|\boldsymbol{x}_{0h'j}, \boldsymbol{x}_{1h'j}, j = 1, 2, ..., J) \geqslant 1 - \frac{\alpha}{2} \tag{9}$$

In addition, we want to explore a slightly more conservative margin $\delta_\lambda = (1-\lambda)\delta$ where $0 < \lambda < 1$. This margin δ_λ can represent the clinically relevant effect that the follow-on generic product should not be worse than the innovator reference product. Examples of λ are 0% (full margin: $\delta_0 = \delta$), 25%, or 50% (half of the margin: $\delta_{0.5} = \delta/2$).

3. Simulation Study

3.1 Simulation Objectives and Plan

In order to characterize the operating characteristics of this Bayesian non-inferiority design for biosimilarity, we will conduct a simulation study that is motivated by our previous example of rheumatoid arthritis. The primary composite efficacy endpoint is the ACR20 at 6 months (or 24 weeks) although ACR50, ACR70, or ACR20 at other time points can be secondary endpoints for generating future hypotheses. There is no safety endpoint in this study and it can be assumed that doses higher than the recommended dose do not create safety concerns.

Table 2: Historical trial on monotherapy of Etanercept (25mg/mL) at 6 months - Moreland *et al.*, 1999

k	Treatment	n_k	$P(ACR20 = 1)$ or $\hat{p}_k$	$\hat{\mu}_{k1}$	$\hat{\mu}_{k2}$	$\hat{\mu}_{k3}$	$\hat{\mu}_{k4}$	$\hat{\mu}_{k5}$	$\hat{\mu}_{k6}$	$\hat{\mu}_{k7}$	σ_k^2
$0h'$	Placebo	80	11%	6%	-7%	2%	-3%	-22%	2%	-207%	1600%
$1h'$	Etanercept (25mg/mL)	78	59%	56%	47%	44%	46%	53%	39%	31%	1600%

Note: For $\hat{\mu}_{k7}$, CRP is used instead of ESR. The variance was reported in Moreland *et al.*, 1997 and was assumed to be 1,600% for the calculation of sample size.

Table 3: Simulation setting for the current non-inferiority biosimilarity trial

Parameter	Values	Description
λ	0, 0.25, 0.5	Sizing factor for non-inferiority margin
$\delta_\lambda = (1-\lambda)\delta$	$\delta_0 = \delta, \delta_{0.25} = 0.75\delta, \delta_{0.5} = 0.5\delta$	Re-sized non-inferiority margin for δ
θ_λ	$\theta_0, \theta_{0.25}, \theta_{0.5}$	Re-sized non-inferiority margin for θ
$\triangle$	-2, 0, 2	Impact of constancy assumption
μ_{1j}, σ_1^2	$\mu_{1j} = \mu_{1h'j} + \triangle, \sigma_{1h'}^2$	Use historical trial on Etanercept (25mg/ML) arm in Table 2
μ_{2j}, σ_2^2	$\mu_{1j} - \theta_\lambda, \sigma_{1h'}^2$	For assessing Bayesian type I error
μ_{2j}, σ_2^2	$\mu_{1j} - \theta_a\ (\theta_a = 0, \theta_\lambda/2), \sigma_{1h'}^2$	For assessing Bayesian power
n	60, 120	Overall trial sample size
R	1, 2	Fixed randomization ratio
p_c	95%, 97.5%	Critical probability
N	determined by simulation	Number of posterior Gibbs samples after 10% burn-in

The ACR20 has seven components and they represent separate categories of symptoms as in Table 1. These components are generally assumed to be independent measures. Therefore, J is equal to 7 such that μ_{k1} and μ_{k2} are at least 20% and at least 3 of $\{\mu_{k3}, ..., \mu_{k7}\}$ are at least 20% where $k = 0h', 1h', 1h, 1, 2$ and $h = 1, 2, ..., H$. The objectives of this simulation study are (1) to assess the type I error in the Bayesian paradigm under the null hypothesis and to compare it with that in the frequentist paradigm, (2) to evaluate the statistical power in the Bayesian paradigm under the alternative hypothesis given overall sample size n and randomization ratio R as well as to compare it with that in the frequentist paradigm, and (3) to characterize the impact of different λ as in δ_λ and p_c on the aforementioned characteristics.

As a real-life motivating example, we conducted a literature search on historical trials on Etanercept. Etanercept is a TNF receptor (p75) fusion protein, linked to the Fc portion of human IgG1. We found five published studies: (1) Moreland *et al.* (1997), (2) Moreland *et al.* (1999), (3) Weinblatt *et al.* (1999), (4) Bathon *et al.* (2000), and (5) Klareskog *et al.* (2004). Among these studies, only one of them (Moreland *et al.*, 1999) was a confirmatory placebo-controlled trial for the monotherapy of Etanercept (25mg/mL) while the other trials studied either combined therapies of Etanercept or lower doses of Etanercept. Etanercept (25mg/mL) was administered subcutaneously twice a week and the primary efficacy endpoint is ACR20 at 6 months (or 24 weeks). This trial led to its FDA approval for RA in 1998. Therefore, $H = 1$, and we will use this historical trial to determine the NI margin as well as including it in the hierarchical bias model $(h' = h)$. Table 2 summarizes the partial result from this historical trial. A positive percent change is interpreted as a reduction in the corresponding symptom component while a negative percent change means an increase. Table 3 describes the simulation setting for the follow-on biological product in the current proposed biosimilarity study.

3.1.1 Simulation Results

Using the estimated means from Table 2 and the simulation setting laid out in Table 3, we simulated patient-level data for the selected historical trial. Using this hypothetical patient-level data, we generated Gibbs samplings on the parameters, $\mu_{0h'j}, \mu_{1h'j}, \sigma^2_{0h'}$, and $\sigma^2_{1h'}$. Using these chains of sampling, we derived the posterior samples of the probability of clinical response ACR20, $p_{0h'}$ and $p_{1h'}$, and hence their difference, $p_{1h'} - p_{0h'}$. The lower bound of the 95% credibility interval is estimated to be 0.4604. Therefore, under different pre-specified sizing factors λ, we can state the different NI margins for subsequent simulation: $\lambda = 0$ will give $\delta_0 = 0.4604$, $\lambda = 0.25$ will give $\delta_{0.25} = 0.3453$, and finally $\lambda = 0.5$ will give $\delta_{0.5} = 0.2302$. We can find their corresponding θ: $\theta_0 = 34.7$, $\theta_{0.25} = 22.2$, and $\theta_{0.5} = 14.0$. Using the same patient-level data, we also calculated the probabilities of clinical response ACR20: $\hat{p}_{0h'} = 0$ and $\hat{p}_{1h'} = 0.5641$ under the frequentist perspective. The estimate for the treatment arm is not far from the one reported in this historical trial (59% in Table 2), but the estimate for the placebo arm is under-estimated (11% in Table 2). The lower bound of the 95% confidence interval is therefore estimated to be 0.4019. The corresponding re-sized NI margins will be 0.4019, 0.3014, and 0.2009. These are somewhat smaller than the corresponding ones estimated in the Bayesian method above. Based on the same simulation plan as described in Table 3, we conducted subsequent simulation using 10,000 simulated identical trials. The same simulated two-arm trial data will be used to determine if the trial is a success separately for the proposed Bayesian method and the standard frequentist method. Figure 2 shows the result of the simulated type I error under both analytical paradigms and Figures 3 and 4 display the result of the simulated statistical powers.

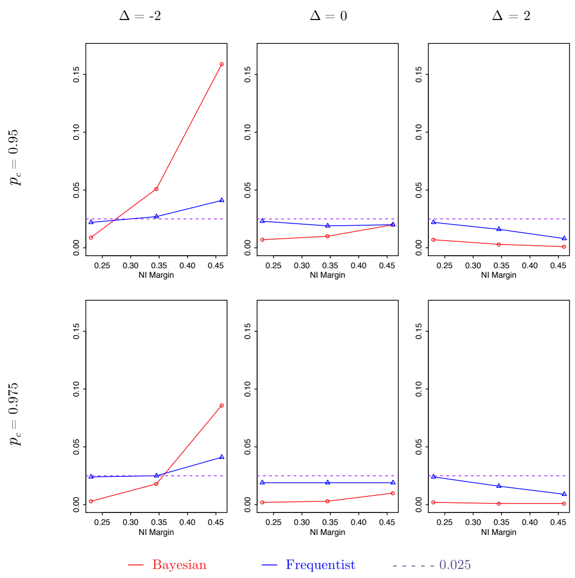

Figure 2: *Plot of type I error against value of* $\triangle = \mu_{1j} - \mu_{1hj}$ *for all* j. *Setting is* $n = 60$, $R = 1$, *and* $p_c = 0.95$

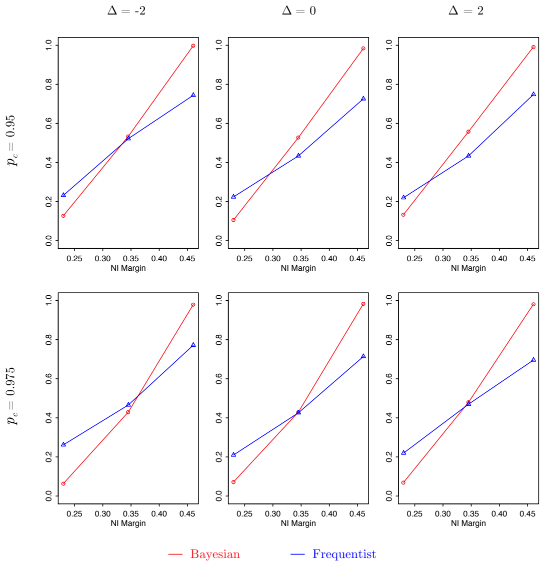

Figure 3: *Plot of type statistical power against value of $\triangle = \mu_{1j} - \mu_{1hj}$ for all j. Setting is $\delta_a = \delta_\lambda/2$, $n = 60$, $R = 1$, and $p_c = 0.95$.*

In Figure 2, we observe general preservation of type I error under 0.025 when $\triangle = \mu_{1j} - \mu_{1hj} = 0$ or 2, but inflated type I error when $\triangle = -2$. That is, when the reference product is performing identically or better in the current biosimilarity trial than in the reference historical trial, the type I error is controlled under the target size. However, if it performs worse in the current trial than in the historical trial, the type I error is inflated, for both the Bayesian and the frequentist methods. However, as λ increases to 0.5 when the NI margin gets closer to 0.2302, the type I error inflation is possible in the frequentist approach but not the proposed Bayesian approach. In fact, the type I error under the Bayesian method is well-controlled under 0.01 even when the reference product is doing worse in the current trial when λ is close to 0.5. Both methods are able to preserve the type I error at 0.025 when $\triangle = 0$, that is when the effect of the reference product is constant in both trials. The inflation of type I error, when reference product is doing worse in the current trial, is possibly due to the larger lower bound of 95% credibility interval in the historical trial as related to the reduced effect size of the reference product comparing to the putative placebo, which does not exist in the current trial. However, as λ increases, the re-sized NI margin narrows, due to the influence of the skeptical prior for ξ_j, the proposed Bayesian method is able to protect the inflation of type I error, even when $\triangle < 0$ but the frequentist approach cannot.

In Figures 3 and 4, we can see that when $\lambda = 0$ and $\triangle = 0, 2$, that is, when the full NI margin is used, the statistical power of the Bayesian method is unanimously higher than that of the frequentist method. As for $\lambda = 0.25$ and $\triangle = 0, 2$, statistical power of the Bayesian method is smaller than that of the frequentist method only when sample size is small as in $n = 60$ and when the alternative is at $\delta_a = \delta_{0.25}/2$. Other than that, the power of the Bayesian method is superior to the frequentist method. As λ increases to 0.5 and

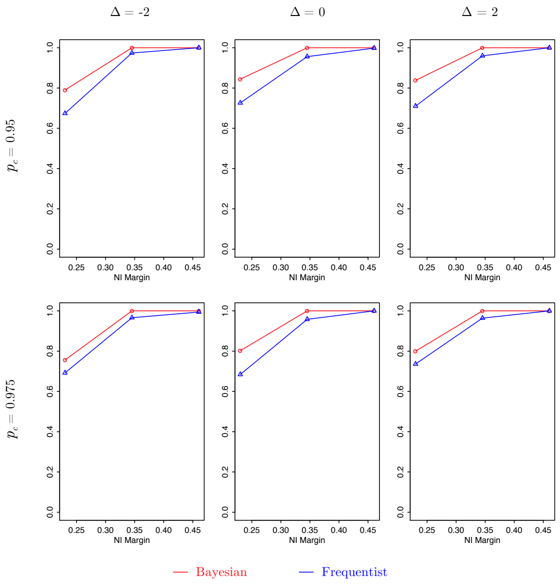

Figure 4: *Plot of statistical power against value of* $\triangle = \mu_{1j} - \mu_{1hj}$ *for all* j*. Setting is* $\delta_a = 0$*,* $n = 60$*,* $R = 1$*, and* $p_c = 0.95$*.*

when $\triangle = 0, 2$, the NI margin narrows down to $\delta_{0.5} = 0.2302$, when $n = 60$ and the alternative is either at $\delta_a = \delta_{0.5}/2 = 0.1151$ or at $\delta_a = 0$, statistical power is very low in both the Bayesian and frequentist approaches with the Bayesian method suffering more loss of power due to the strong influence of the skeptical prior on ξ_j within the smaller margin. However, an increase in sample size promises a much better improvement in statistical power when $\delta_a = 0$ than the improvement in frequentist power. This is mainly due to the increasing influence of the data over the skeptical null prior, resulting in improved Bayesian power.

4. Summary and Discussion

In this paper, we have presented a Bayesian method to assess biosimilarity between a licensed reference biological product and a generic follow-on (also known as a subsequent-entry) biological product. This approach adopts a non-inferiority testing framework that connects the current trial of biosimilarity to historical trials of the reference product. The proposed Bayesian analytical approach recognizes that the reference product was approved for license in the past and that information in these historical trials can be meaningfully incorporated in the analysis of the current trial. However, due to changing clinical practices and improvement in the overall delivery of care over time, the effect of a medicinal product may not be always constant. This is, in the context of a non-inferiority clinical trial, sometimes known as the constancy assumption, the historical difference between the original product and placebo is assumed to hold in the current setting of the new trial if a placebo is in place (D'Agostino, Massaro, and Sullivan, 2003). Therefore, we presented the hierarchical model to incorporate historical trials while accounting for the potential lack of biosimilarity via a bias parameter. In this model, non-informative priors are elicited for most parameters except for the bias parameter which assumes a skeptical prior with expectation centered on the null hypothesis. As most biological products are meant to treat illnesses with improvement in multiple endpoints, we illustrate the application of this method to studying rheumatoid arthritis that uses a composite efficacy endpoint known as ACR20.

Simulation studies have demonstrated that the Bayesian method usually has type I error preserved under the α-level of 0.025, comparable to a typical level assumed in a one-sided non-inferiority trial. This is made possible with the placement of the skeptical prior on the bias parameter, even when a more relaxed critical probability $p_c = 0.95$ is used. When the reference product performs worse in the current trial, due to potential violation of the constancy assumption, the NI margin that is based on its historical trial appears to be wider, thus inflating its type I error. Both Bayesian and frequentist methods have no immunity to this inflation, however, the Bayesian method is able to cancel out this inflation by tapping into the influence of the skeptical null prior as NI margin narrows, therefore offering some protection even when constancy assumption is slightly violated in the negative direction. It is important to emphasize that this type I error is an error rate conditional on the outcomes of the historical trial selected. Under this hierarchical model, we presume in (5) that both μ_{1j} and μ_{1hj} come from the same underlying distribution, therefore the difference $\triangle = \mu_{1j} - \mu_{1hj}$ follows the normal distribution, $N(0, 2\sigma_{1b}^2)$. Another way to look at the type I error is the average type I error rate over all possible values of $\triangle$. Further simulation can be useful in characterizing this average type I error over all possible trial performance for the reference product in historical and current trials. It is important that, prior to the design of the biosimilarity trial, a thorough literature search should be made to assess if the effect of the reference product is consistent in the historical trials and if the design and conduct of these studies are not too dissimilar. If such large variability in estimation is observed, sources of this inconsistency should be investigated.

As for statistical power, it somewhat suffers when NI margin is small. However, as sample size increases from $n = 60$ to 120 under smaller margins and as the follow-on product is truly biosimilar to the reference product, we expect the Bayesian statistical power to outperform the frequentist approach. In addition, it will be interesting to explore a Bayesian two-stage adaptive design using predictive probability as an interim stopping criterion. It is possible to further reduce the expected sample size especially in cases when a follow-on product is biosimilar to the reference product without compromising its statistical power.

Another possibility of using hierarchical modeling is that we may be able to include other historical trials which perhaps studied different doses of the reference product or were conducted under systematically different trial-specific circumstances. If such characteristics can be assumed to be linearly related to the efficacy parameters, their inclusion into the model may help increase the precision of the estimation, and hence the inference.

In this paper, we have illustrated the method using a composite endpoint that has several separate endpoints combined into a single one. When we directly model the component endpoints, it is likely that instead of the global null hypothesis, some of the component endpoints may have inferior means such that for some j, $\mu_{2j} \leqslant \mu_{1j} - \theta$ but not the others, and this trial can still claim success based on the predictive or posterior probability. Composite endpoint may present different null configurations which may warrant further study. In our example, we have only presented the global null configuration using θ as the non-inferiority margin across all component endpoints. In other cases, a single endpoint or multiple endpoints are used to establish efficacy. For example, for studying psoriasis, a common chronic inflammatory skin disease characterized by thick red flaky patches called scales, there are two major endpoints: proportion of subjects who achieved at least 75% reduction in PASI score (PASI75) and treatment success on the Physician's Global Assessment (PGA). This Bayesian hierarchical bias approach can still be similarly applied and final inference may be based on the joint posterior probabilities that these endpoints are greater than their respective non-inferiority margins.

REFERENCES

Bathon, J.M., Martin, R.W., Fleischmann, R.M. et al. (2000). ”A comparison of etanercept and methotrexate in patients with early rheumatoid arthritis.” *The New England Journal of Medicine*, Vol. 343, No. 22, pp. 1586-1593.

D'Agostino, R.B., Massaro, J.M., Sullivan, L.M. (2003). ”Non-inferiority trials: design concepts and issues - encounters of academic consultants in statistics.” *Statistics in Medicine*, Vol. 22, 169-186.

FDA. *Scientific Considerations in Demonstrating Biosimilarity to a Reference Product.* The United States Food and Drug Administration: Rockville, MD, 2012.

Felson D.T., Anderson J.J., Boers M., et al. (1993). ”The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials.” *Arthritis Rheumatology*, Vol. 36:729-40.

Gamalo, M.A., Wu, R., Tiwari, R.C. (2012). ”Bayesian approach to non-inferiority trials for normal means.” *Statistical Methods in Medical Research*, doi: 10.1177/0962280212448723.

Gamalo, M.A., Tiwari, R.C., LaVange, L.M. (2013). ”Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products.” *Pharmaceutical Statistics*, doi: 10.1002/pst. 1588.

Kang, S.H., Chow, S.C. (2012). ”Statistical assessment of biosimilarity based on relative distance between follow-on biologics.” *Statistics in Medicine*, 32: 328-392.

Klareskog, L., van der Heijde, D., de Jager, J.P. et al. (2004). ”Therapeutic effect of the combination of etanercept and methotrexate compared with each treatment alone in patients with rheumatoid arthritis: double-blind randomised controlled trial.” *The Lancet*, Vol. 363, pp. 675-681.

Lin, J.R., Chow, S.C., Chang, C.H., Lin, Y.C., Liu, J.P. (2012). ”Application of the parallel line assay to assessment of biosimilar products based on binary endpoints.” *Statistics in Medicine*, 32: 449-461.

Moreland, L.W., Baumgartner, S.W., Schiff, M.H. et al. (1997). ”Treatment of rheumatoid arthritis with a recombinant human tumor necrosis factor receptor (p75)-Fc fusion protein.” *The New England Journal of Medicine*, Vol. 337, No. 3, pp. 141-147.

Moreland, L.W., Schiff, M. H., Baumgartner, S.W. et al. (1999). ”Etanercept therapy in rheumatoid arthritis.” *Annals of Internal Medicine*, Vol. 130:478-486.

Offen, W., Chuang-Stein C., Dmitriendko, A., et al. (2007). ”Multiple co-primary endpoints: medical and statistical solutions. A report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America.” *Drug Information Journal*, Vol. 41, 00. pp. 31-46.

Pocock, S.J. (1976). ”The combination of randomized and historical controls in clinical trials.” *Journal of Chronic Diseases.* Vol 29, 175-188.

Erickson, W. P., McDonald, T. L., Gerow, K. G., Howlin, S., and Kern, J. W. (2001), “Statistical Issues in Resource Selection Studies With Radio-marked Animals,” in *Radio Telemetry and Animal Populations*, eds. J. Millispaugh and J. Marzluff, California: Academic Press, pp. 209–242.

Weinblatt, M.E., Kremer, J.M., Bankhurst, A.D. et al. (1999). ”A trial of etanercept, a recombinant tumor necrosis factor receptor: Fc fusion protein, in patients with rheumatoid arthritis receiving methotrexate.” *The New England Journal of Medicine*, Vol. 340, No. 4, pp. 253-259.